Stereo Vision Motion Detection from a Moving Platform

Gary Overett[1]
gary@syseng.anu.edu.au

David Austin[1,2]
d.austin@computer.org

[1]Robotic Systems Lab, RSISE
Australian National University,
ACT 0200, Australia

[2]National ICT Australia,
Locked Bag 8001,
Canberra, ACT 2601

Abstract

In order for robots to coexist with humans it is important that they are able to see and track them. Simple tasks like moving to avoid a passing human in the corridor require that the robot is able to estimate the position and velocity of the passer. Several systems for tracking people have been presented in the past, however few attempt tracking using vision on a moving platform. In the field of mobile robotics it is important that the methods used in tracking are robust to egomotion. In this paper we present preliminary results of a system which uses basic stereo vision to segment and track moving objects. The system requires the use of both the three dimensional stereo disparity information as well as optical flow to segment and track people. Experimental results showing the removal of optical flow due to egomotion are presented as well as a method for measuring the change in depth over time after compensating for the robots motion. We also present a useful method for judging the quality of the gathered visual data. Finally the capability to track moving people using a Kalman filter is demonstrated.

1 Introduction

As robots move from common use in highly engineered environments to more general settings they will require the skills to cohabitate with humans. While old factory robots may have been able to exist in separate areas to humans the robots of the future will have to be able to monitor and adjust to the unpredictable behaviours of the humans with which they share space.

Consider the example of an autonomous office cleaning robot. It is not acceptable for it to ignore people in the environment, while they move around the robot and avoid it. Rather, it is highly desirable that the robot should be able to react to the presence of humans appropriately. It makes sense, for example, for the robot to move to the side in response to people passing by. As with people in the office, the robot will have to choose which side is best to move to in order to least inconvenience the passer. If the robot is able to detect the path of the incoming person it will be able to move out of the way. One of the most simple human-robot interactions would be simple corridor avoidance, where the presence of an approaching human is met by a move to the side by the robot. In other cases the robot's express purpose may be to interact with humans, for example, as a tour guide robot [Saragih *et al.*, 2004]. In this case the robot, like a real tour guide, will require information about the whereabouts of its subject audience relative to itself. In addition to this it will be useful for such a robot to be aware of the movements of its audience. Even in the case where the robot has a complete map of its surroundings and the ability to navigate robustly in its environment it is still important for the robot to be able to follow the movements of the people around it and react as required. Yet another useful robot behaviour is following in order to build a map. In this case the robot is required to track a person and follow them around the scene.

While several studies have been performed in order to track humans using various vision systems and sensors such as laser scanners [Fod *et al.*, 2002], few present methods which are robust to the movement of the observing platform. Many researchers assume a static system in order to use background subtraction algorithms, such as those used in [McKenna *et al.*, 2000] and [Bahadori and Iocchi, 2003]. While the assumption of static cameras works well for tracking applications such as surveillance and intelligent monitoring, it produces systems which cannot work on moving platforms such as mobile robots. Some very effective tracking systems such as [Sogo *et al.*, 2000] and [Mittal and Davis, 2002] rely on multiple views from around the room. While their results are impressive, they are not practical for use in most mobile robotics applications. Almost no research

has been conducted to explore the possibility of tracking aboard mobile robots. Thus, the major aim of this project is to develop a system which tracks people using the CeDAR stereo vision system [Sutherland *et al.*, 2000] mounted on top of a mobile XR4000 robot (see figure 1).

2 Experimental Platform

This project relies on the the use of the ANU's XR4000 mobile robot. In short this robot provides the mobile platform from which we are able to take our stereo images. The XR4000 can move freely and rotate. The robot also provides the odometric data, which is used to remove the egomotion of the robot from the optical flow.

The CeDAR active head provides the stereo camera platform and contains two S-Video camera's which are able to provide live camera data and produce video data. Currently we are working to collect better odometric data, which more accurately lines up with the camera images produced. At the moment, we rely on somewhat manual calculation of the odometric data pertaining to a specific set of camera images.

The XR4000 robot also has a front mounted SICK Laser Scanner which provides two dimensional depth ranging. This is useful as a means of testing the accuracy of basic depth measurements calculated from the stereo vision system.

3 Approach

The system presented first analyses the flow and depth information from the stereo cameras and then attempts to calculate the image flow when the egomotion of the robot is accounted for. For example, when the robot is moving left the optical flow in the images moves to the right. We calculate the expected flow, for the associated depth, in the image and subtract that from the observed flow. Thus only the parts of the image which are moving relative to world coordinates are tracked as moving, while the stationary background is removed.

Optical flow is notoriously noisy and we need a method to judge the validity of the flow information gathered. To this end, we present a quality measure which in most cases will give a good indication as to the quality of both the disparity and flow information used. Then by paying close attention to this information we attempt to segment moving objects. This segmentation result will ultimately be combined back with the depth and flow information to track the subjects in the scene.

A system diagram showing the proposed layout of this method can be seen in figure 2.

3.1 Processing Speed

In this research we have stepped over the complications of real-time processing to achieve the main result quicker. All the work shown here is done using video sequences

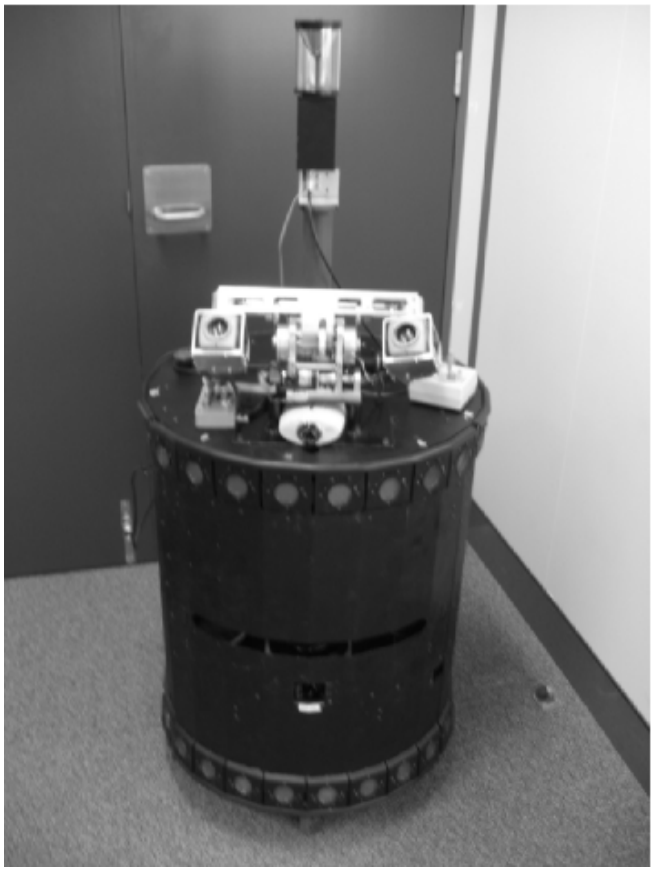

Figure 1: The ANU XR4000 with CeDAR Active Head

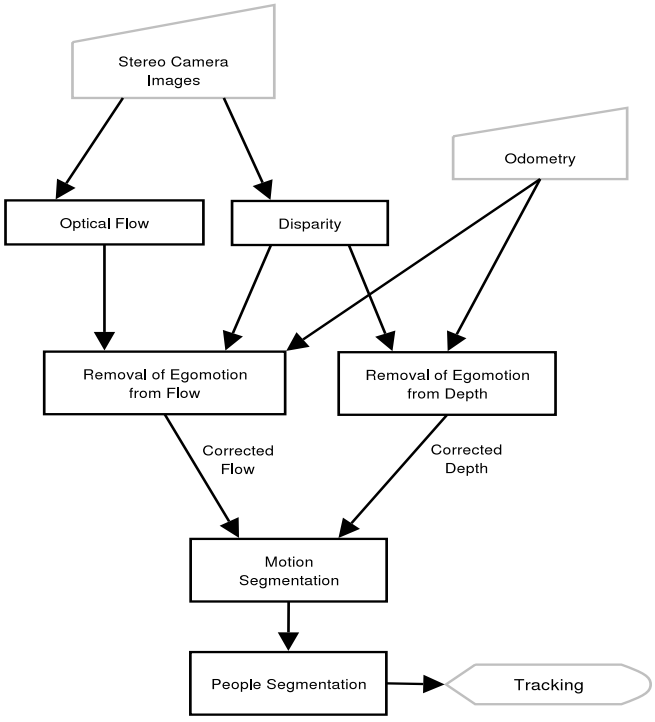

Figure 2: System Diagram

and could not be done in real-time using the current system. This allows us the opportunity to explore more in depth processing techniques for each set of frames in the video sequences. Most of the processing time in this research is taken up by the processing of disparity and flow in the images, which some researchers are starting to optimise to the degree that would allow our system to be implemented in real-time. Section 8.5 discusses the potential of the system to be produced in real-time using some existing optimisation techniques.

4 Sum of All Differences Correlation for Stereo Depth Analysis

Disparity is the key element of depth perception using stereo vision. Alternately, disparity can be calculated in many ways with differing results for the speed, accuracy and noise level. We use an adaption of the SAD method as used by [Thompson, 2002] and [Watman *et al.*, 2004] for the correlation used to measure disparity.

The SAD method is employed to compare parts of the primary camera image to find the corresponding section in the secondary camera image. The camera poses are known, allowing the search window to be restricted to a reasonable size and allowing a greater field of view for the cameras by maximising the width of the valid region for disparity analysis. For each pixel in the search window the correlation is found using

To speed things up we add a resolution factor R, which allows the correlation to be performed using only every R^{th} pixel in the images P and I. This scaling factor is equivalent to reducing the resolution of the primary and secondary images except that the disparity is still calculated to within 1 pixel of the original image. The scaling factor R allows us to use a large template size of 32 by 32 pixels without necessarily having to do 32^2 pixel comparisons. The template size used was found to produce the best measurements for both disparity and optical flow for typical images. Large templates tend to 'blur' disparity and flow measurements around the edges of objects because they capture both background and foreground sections in the one template. For smaller templates we tend not to capture sufficient structure in a template to find an unambiguous match in the secondary image. The template size chosen catches sufficient structure at the scale of moving people in the image and produces accurate results. While significant noise does occur at the edges of objects in the scene this is usually filtered by the quality measure, which is presented in detail later in this paper (see section 5). The disparity search window chosen is 16 by 128 pixels. These search window dimensions allow for relatively large disparities to be searched (up to 128 pixels), which in turn allows the subjects in the scene to be tracked over a depth range of infinity to about 1.4 meters. The 16 pixels of height in the search window allow for some inaccuracies in the line up of the camera mountings and aid the calculation of the quality measure used.

$$SAD_{(x,y)} = \sum_{i=1}^{M/R} \sum_{j=1}^{N/R} |I_{(x+Ri,y+Rj)} - P_{(x+Ri,y+Rj)}| \quad (1)$$

where the primary camera image template P is correlated with an image I from the secondary camera image chosen within the search window.

The resulting disparity is found by searching for the sub-image in the secondary camera search window, which minimises the SAD when compared to the template in the primary camera image. Then the disparity is the difference in the x coordinates of the image points from both the primary template and secondary camera sub-image. An example of the disparity can be found in figure 6.

5 Quality Estimates for SAD Correlations

Since depth and flow analysis tend to be noisy it is important that we have some concept of the accuracy of our data. To produce a general quality measure we must first characterise the nature of a high quality correlation from the SAD process. In the case of [Thompson, 2002] and [Watman *et al.*, 2004] a good quality template T was characterised roughly as *a template image T such that no similar images can be found in some neighbourhood N of T*, and thus the similarity found in N was the metric of quality. In the case of disparity it does not make sense to gauge quality by comparing the template P with the neighbourhood in the primary camera image, but rather to compare it with its associated search window in the secondary camera image. However, the success of the aforementioned quality measure suggests the nature of a good quality measure for disparity information.

In our case we characterise a quality disparity match between the primary camera template P and its match m in the secondary set of images S in the search window as *A template P in the primary image which has few matches $M \subset S$ which correlate nearly as well as the best match m.*

This implies the following mathematical definition of quality:

$$Q = \frac{1}{\sum_{i=m}^{m+n} H(i)} \quad (2)$$

where m is the minimum SAD value in the histogram H of the frequency of different correlations in the search window, n is the distance to sum along the histogram and $H(i)$ is the histogram value at i.

An important property of this quality measure is that even if the SAD is high for the best match m found[1], the value of Q may still be high. This is a desirable property because poor surfaces such as walls tend to yield very low SAD values over the full search window while highly structured sections may not match as well. This can be observed in the histograms shown in figure 4. In A we observe a minimum sad value of 3 (very high correlation) for the section of white wall as shown in figure 3. That is, there are a large number of well correlated sections of white wall to match the template A of white wall. However, for the very structured template C we observe a minimum sad of just 18, but will find a high quality Q because the solution sub-image is significantly better than the other matches found in the search window.

Returning to equation 2, we see that Q is simply the sum of the first n elements of the histograms shown in figure 4. For our purposes, we choose $n = 10$, which produces good results.

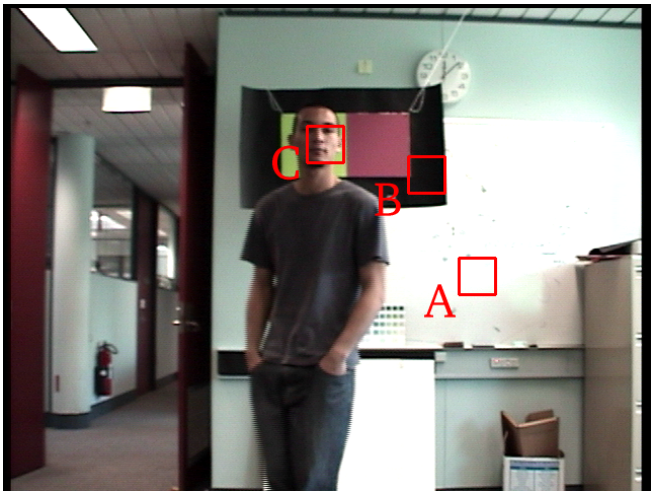

Figure 3: Camera frame containing templates A,B and C, which are low, medium and high quality templates respectively.

6 Removal of Egomotion from Optical Flow

Previous researchers have used optical flow from images taken aboard a mobile robot to calculate visual odometry [McCarthy and Barnes, 2004] [Björkman and Eklundhs, 1999]. In this research we turn the technique around, taking the robot odometry as given (by motor odometry), and attempt to use this to remove the optical flow due to egomotion. This results in a means to judge whether a particular part of the image is moving.

[1]High SAD values correspond to a lower correlation

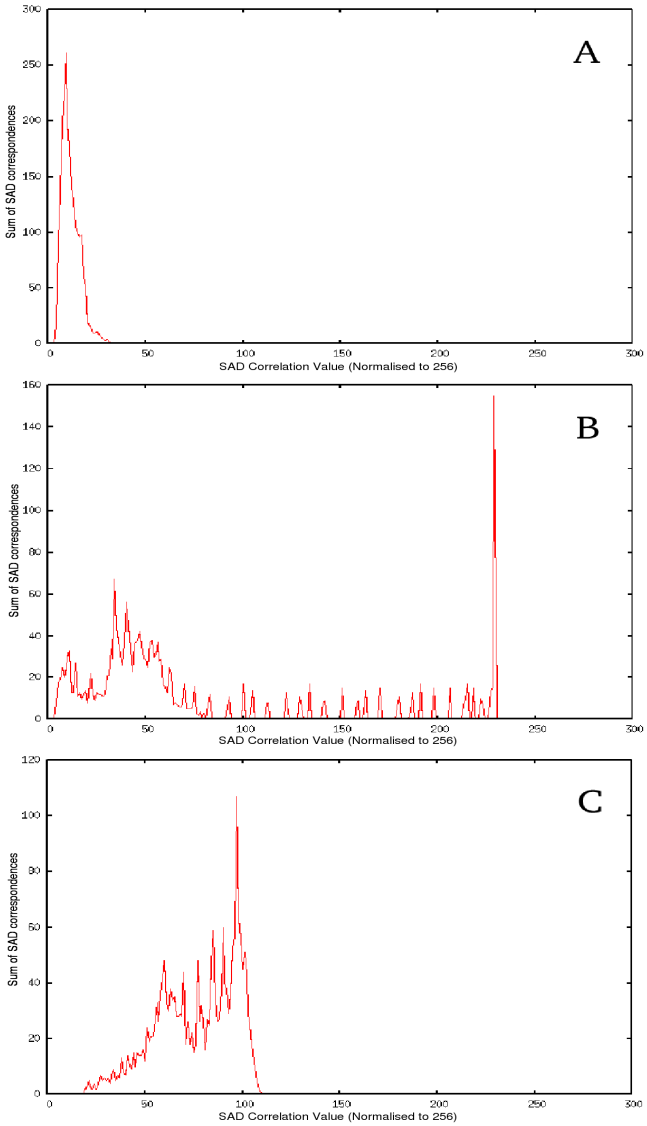

Figure 4: SAD Histogram data for matches in the disparity search window. The 3 histograms are for the low (A), medium (B) and high (C) quality templates shown in figure 3. The quality factor Q will be the sum of the first n non-zero elements of the histogram array shown above and denoted by equation 2.

Before removing the egomotion from the flow we must have already calculated the depth from the disparities found. In fact, we calculate the coordinates for each point whose disparity is calculated. Thus this information is already available to us when removing the egomotion.

Thus we know the points P_n and the egomotion E_n.

$$P_n = \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix}, E_n = \begin{bmatrix} e_{x_n} \\ e_{y_n} \\ e_{a_n} \end{bmatrix}$$

Note that e_{a_n} is the angular egomotion.

Thus we can find the new expected scene points for this motion as:

$$P_{n+1} = \begin{bmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} x_n - e_{x_n} \\ y_{n+1} - e_{y_n} \\ z_n \end{bmatrix} \quad (3)$$

Next we find the horizontal and vertical change in angle, relative to the primary camera, for the scene points provided:

$$\theta_x = \tan^{-1}(\frac{x_{n+1}}{y_{n+1}}) - \tan^{-1}(\frac{x_n}{y_n}) \quad (4)$$

$$\theta_y = \tan^{-1}(\frac{z_{n+1}}{\sqrt{x_{n+1}^2 + y_{n+1}^2}}) - \tan^{-1}(\frac{z_n}{\sqrt{x_n^2 + y_n^2}}) \quad (5)$$

Next we calculate the new flow vector G_n from the old flow vector

$$F_n = \begin{bmatrix} f_{x_n} \\ f_{y_n} \end{bmatrix}$$

as:

$$G_n = F_n + \begin{bmatrix} (\theta_x + e_{a_n})\phi_{cam} \\ \theta_y \phi_{cam} \end{bmatrix} \quad (6)$$

where ϕ is the camera parameter denoting the pixels per radian for the cameras used

Thus we find G_n the flow vector, corrected for the egomotion of the robot. The practical results from this work can be found in section 8.2.

7 Removal of Egomotion from Depth Measurements

When people move toward and away from the robot along the optical axis the flow due to their motion is minimal and therefore cannot be segmented. To cover this case we implement a second method for removing the robots egomotion from depth so as to segment motion based on the perceived change in depth. This method is somewhat problematic because the depth information is particularly noisy.

To find the new scene points P_{n+1} between frames we are able to apply the same technique as used in Equation 3. Producing the scene points from the new is difficult because we want to know the expected depth of the scene along specific directions (those seen by the camera from the new view). Ultimately the robots motion may mean that we are looking into a direction along which we have no previous scene points. That is, no previous information. To produce these estimates we first attempt to find a scene point from the set P_n in approximately the same direction. Fortunately, this method is successful method most of the time because the robot tends not to move too far over the small time steps.

8 Results

8.1 Quality and Depth Measures

Perhaps the best demonstration of the success of the quality measure in gauging the quality of a measured disparity is to compare the quality Q to the standard deviation ω_{depth} of the depth over several frames while the robot is kept stationary. For templates where Q is high, we find that ω_{depth} is very low. This shows that quality templates can be relied upon to produce good quality depth measurements (see figure 5). One unfortunate fact is that most of the data in the image is in fact of a quality that produces noisy depth results. Fortunately, most people being tracked can be expected to naturally have sufficiently detailed image structure to allow accurate results. This can be seen in figure 6, where the subject in the images has a large number of high quality matches in the stereo images.

Figure 6 shows the disparity image and quality image with the original stereo images shown alongside. In figure 7 we show the depth estimates from figure 6 projected onto a two dimensional map of the room. This image is produced by adding the quality Q associated with each disparity calculation to the (x, y) coordinates in the robot map, which are associated with this calculation. While the background does not exhibit much structure in two dimensions, we can clearly see the position of the person shown in figure 6.

8.2 Removal of Egomotion from Optical Flow

We found that the removal of egomotion from optical flow calculations worked well for reasonably detailed regions of the image. Figures, 8 and 9 show the removal of the egomotion for some more detailed images taken inside the lab.

For less detailed scenes such as an ordinary office corridor we see that the process doesn't work as well. In figure 10 we observe that the flow in the x direction is well corrected but that the original flow is noisy in the y

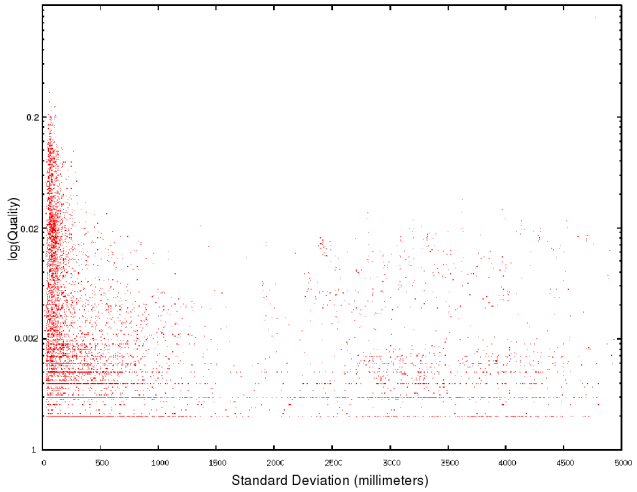

Figure 5: Quality vs Standard Deviation of Depth Estimates over 8 Static Frames. From this we can see that high quality templates, with $Q > 0.02$ produce very stable depth measurements. Depth measurements are reliable for $Q > 0.002$, but may contain a few very poor values.

direction due to the fact that most detail in the images is in the form of vertical lines from door posts etc.

Of course for angular egomotion we find the the corrected flow is almost always good. This is because the angular flow is independent of depth in the image and thus has only to cope with the noise in the original flow. This is shown in figure 11.

Overall, we are able to remove the part of the flow contributed by the robot egomotion enough to feed the segmentation process to follow. Generally, we would like to be able to segment and track people even if they move slower than the robot with respect to world coordinates, and this appears to be the case.

8.3 Segmentation of Moving Regions

The removal of the egomotion from depth information is, in itself, very successful. However, because depth information is so noisy we do not achieve very accurate results when attempting to determine motion from this information. The major difficulty is that between two time steps, a given scene point may be measured as being at different depths even if it has not moved. For this reason, real movement is hard to separate from noise.

Figure 13 shows the change in depth information for the image pair shown in Figure 12. Figure 14 shows the resulting segmentation from depth.

Next we attempt to segment moving objects by masking out regions which are not of a high enough quality (that is, we expect them to be noisy) and which do not exhibit large enough optical flow. Here we observe that

Figure 6: Disparity and Quality Using the Left Image as the Primary Image. We note that people generally exhibit enough image structure to produces reasonable quality ratings.

Figure 7: Two Dimensional Map of the Robot View *Note: brightness corresponds to the quality of the depth estimate and is shown in log(Q) form to show up low quality results as well.* The dark blob corresponds to the person shown in figure 6 and the mark at the bottom of the figure marks the position of the CeDAR.

Figure 8: Optical flow removal for a forward move of 15cm. The blue lines are the original flow while the white lines are the corrected flow. Note that low quality flows have been removed.

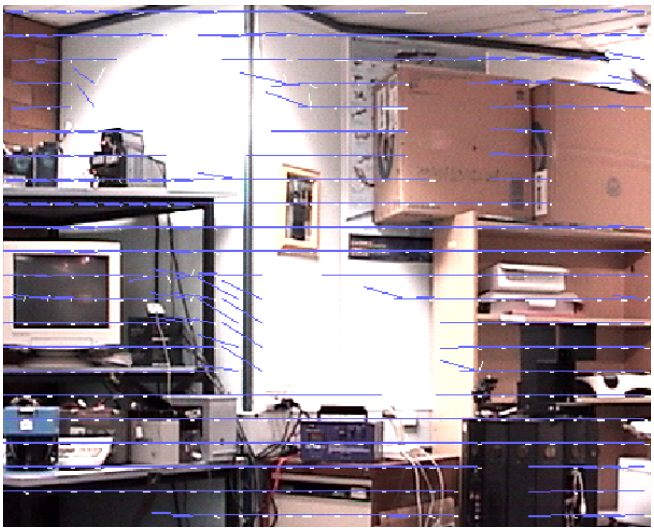

Figure 9: Optical flow removal for a left move of 15cm. The blue lines are the original flow while the white lines are the corrected flow. Note that low quality flows have been removed.

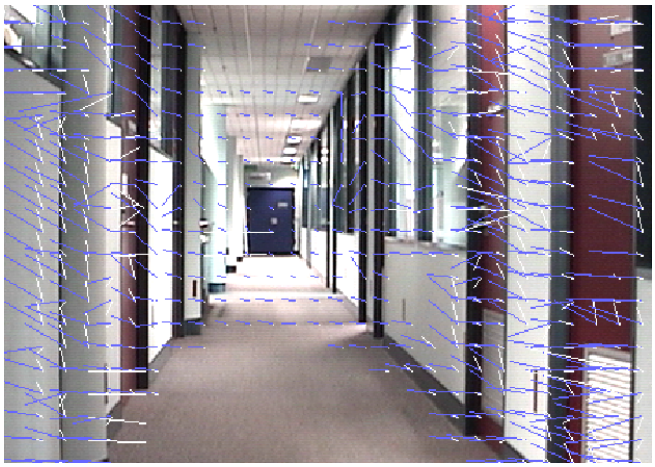

Figure 10: Optical flow removal for a left move of 15cm. The blue lines are the original flow while the white lines are the corrected flow. Note that low quality flows have been removed. In this image we observe that the flow in the x direction is removed well but not in the y direction. This is attributed to the abundance of vertical door posts etc, but not much horizontal structure.

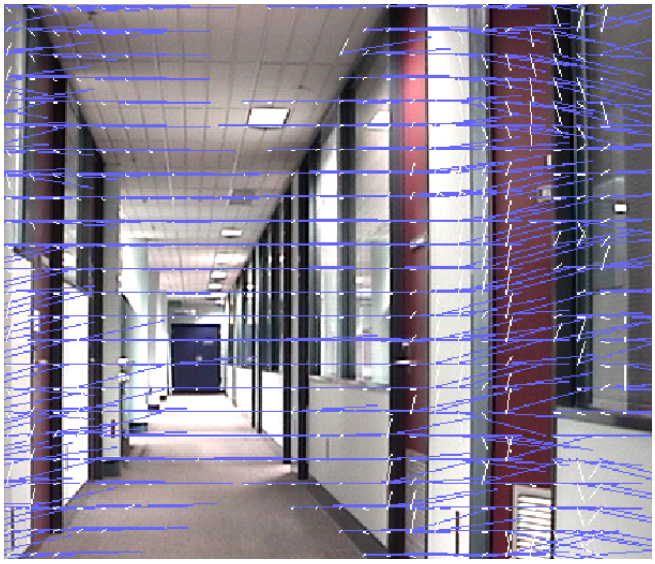

Figure 11: Optical flow removal for a rotation of 4.5 degrees. The blue lines are the original flow while the white lines are the corrected flow. In this image we observe that rotations in general will be removed from the egomotion very effectively. This is because rotations are independent of depth measurements.

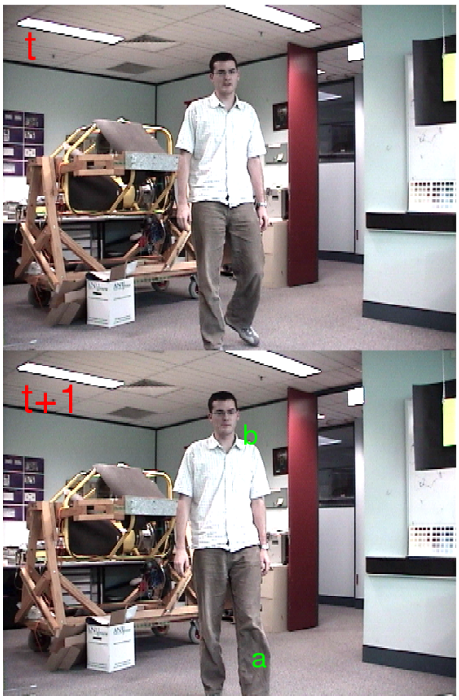

Figure 12: Images used in to find the change in depth information shown in Figure 13

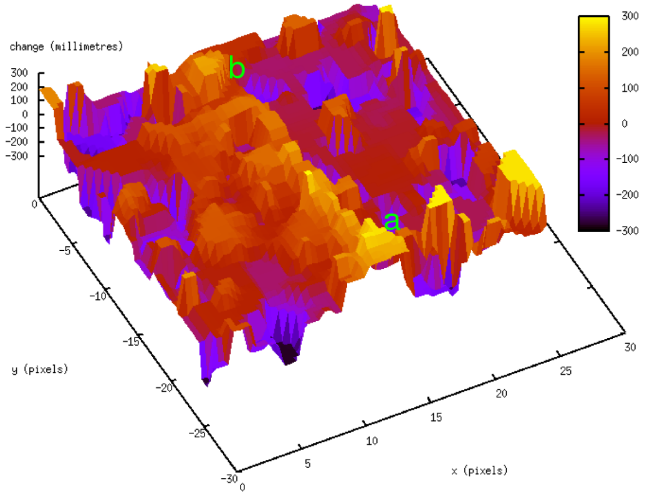

Figure 13: Change in depth information calculated over two frames. In this image the robot moved forward about 10cm. The region marked (a) and (b) show the corresponding points is the Figure 12. The high noise level is easily observable.

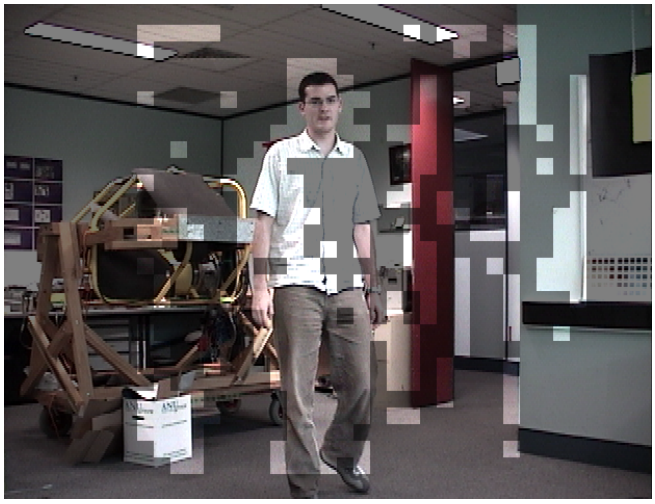

Figure 14: Segmentation of the scene as shown in Figure 13.

if the objects in the scene do not move enough to have a significant optical flow due to their own motion, then we will not be able to segment them because they will be lost in the noise level. We found that under normal robot movements the noise level for most of the image is kept below 7 pixels. This means, that to segment motion in we simply remove regions with less than 7 pixels of flow.

Figure 15 shows an image with a moving subject, before and after the flow from egomotion has been removed.

Figures 16 and 17 show the segmented images for two moving scenes. Figure 16 is the segmented image belonging to the images shown in 15. Thus we are able to see that the flow is easily segmented for typical camera images, and is well above the noise level. This should provide a strong beginning for the process of tracking motion as this project continues.

8.4 Tracking

The final stages of our system uses a series of methods to clean up the final segmented information and find the centroids of the people moving in the scene. We then use a simple Kalman filter [Welch and Bishop, 2004] [Kleeman, 1996] [Tomasi, 2000] to track the subjects in the scene. This gives us some prediction of their future movements and allows us to track even when people fail to be segmented accurately from the raw data. An example of this is seen in Figure 18.

8.5 Computational Speed

Table 1 shows the various computational times for the system. The flow pictures shown in this paper were created using a step of 16 for both flow and disparity. While these results do not allow for real-time processing of the

Figure 15: Optical flow removal for a scene with a moving subject. A shows the originally calculated flow and B shows the corrected flow once ego motion is accounted for. In this scene the robot moved 17cm left while the subject followed the robots motion.

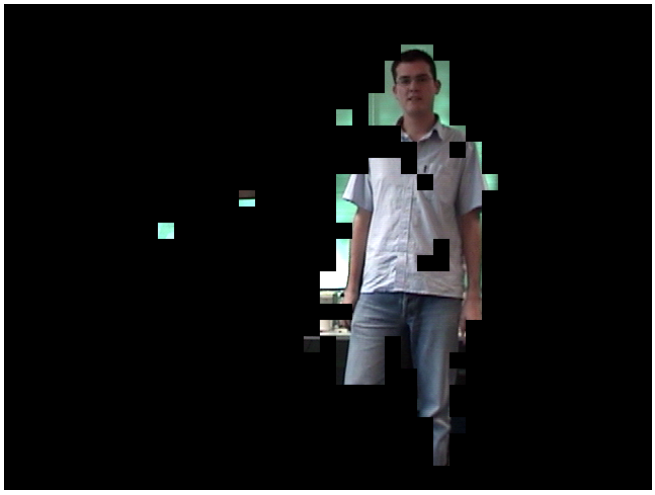

Figure 16: Segmented subject moving alongside the robot as shown in figure 15. Note that only a small amount of noise is observed.

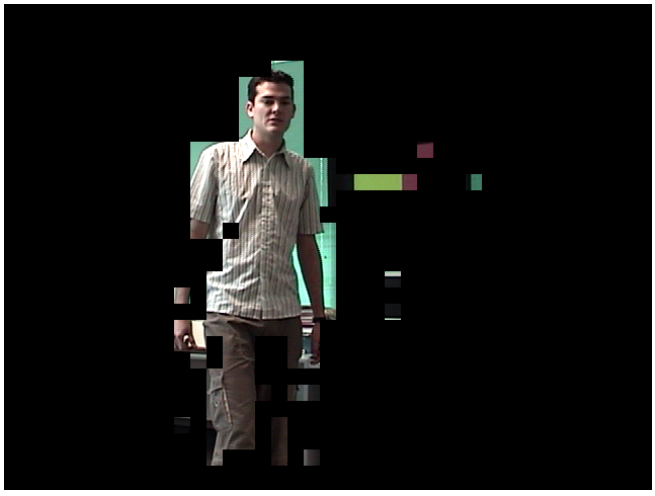

Figure 17: Segmented subject moving in a scene while the robot rotates.

Calculation	Step	R	Speed (sec)
Flow	16	2	≈7
Flow	16	4	≈1.8
Disparity	16	2	≈3.5
Disparity	16	4	≈1
All Other Work	-	-	negligible
Total	16	2	≈10.5
Total	16	4	≈2.8

Table 1: Table of Computational Times using Various Stepping Distances for Analysing the Images
Note: this is only using a single camera as the primary image. If we use both cameras as the primary image, one after the other, it will double the computation time.

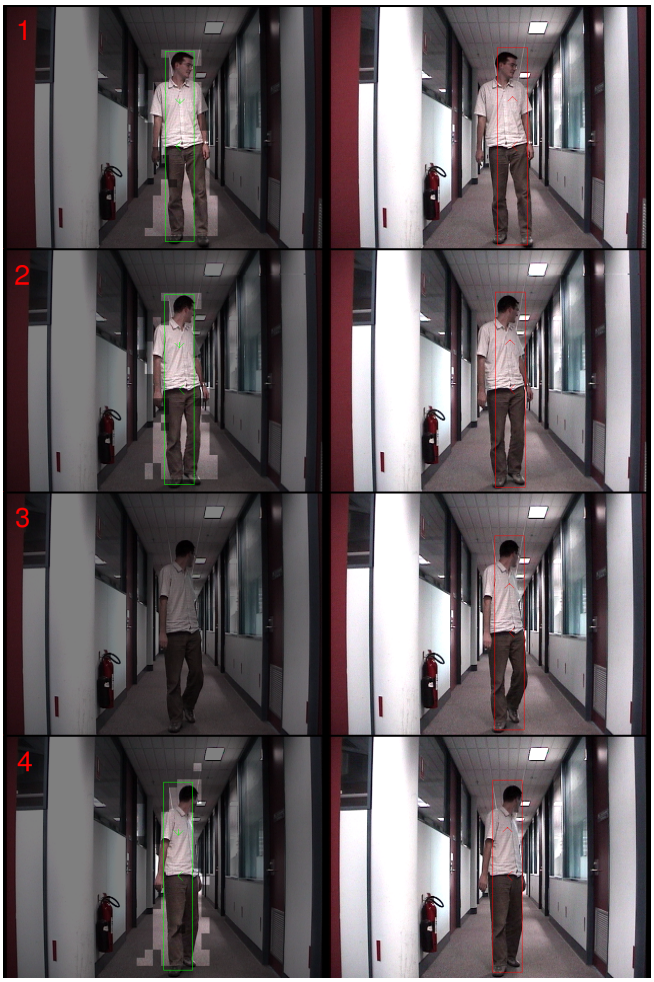

Figure 18: Tracking Sequence showing a Bridged Gap. The left frames are the original instantaneous segmentation and the right images are the Kalman filter tracking results.

data, they do show that such algorithms could be done in real-time in the future.[2]

It is believed that future versions of this system could be done using optimisations such as those used by [Björkman and Eklundhs, 1999] and VLSI stereo processing as in [Porr *et al.*, 2002].

A basic analysis examining the feasibility of real-time processing suggests that this is possible. The analysis suggested that a speed of about 0.08 seconds per frame would be possible. Such a speedup would rely on a series of optimisations such as the use of MMX techniques as seen in [Fletcher *et al.*, 2003] and [Björkman and Eklundhs, 1999], which produce a 4 fold speedup. Major gains can also be made by speeding up the frame rate of the original footage to allow for smaller search windows to be used. This introduces noise though this can be filtered out with temporal filtering techniques like those used in [McCarthy and Barnes, 2004].

9 Conclusion

This system demonstrates the success of a general method for the removal of egomotion from both optical flow and disparity. This leads to a successful means to segment the moving regions of the scene when objects move in any direction. While both optical flow and disparity provide rich information about motion in the scene the segmentation from motion compensated optical flow is far less prone to noise.

To help protect the system from noise we have demonstrated the success of a general quality measure, which provides a strong indication of the expected accuracy for different parts of the scene.

The ability to use this information in tracking is also shown with a high level of accuracy and robustness. The final system places few restrictions on the movement of

[2]These results were taken using an AMD Athlon XP 2700+

the robot or the people moving in the scene as the tracking works best under normal conditions for an indoor environment. This provides researchers with a robust method for tracking people using stereo vision from mobile robots.

The optimisations used are successful at reducing the speed of computation by a factor of 4 and bring the system within reach of real-time operation. Most of the processing time is used by the disparity and flow estimations which may in the future be moved to specialised processing modules. Thus, future systems may well demonstrate real-time tracking with similar accuracy.

Overall the system has surpassed our expectations and shows great promise for helping robots advance towards use in less constrained environments.

References

[Bahadori and Iocchi, 2003] S Bahadori and Luca Iocchi. A stereo vision system for 3d reconstruction and semi-automatic surveillance of museum areas. *Eight National Congress of Associazione Italiana per l'Intelligenza Artificiale*, 2003.

[Björkman and Eklundhs, 1999] M. Björkman and J-O. Eklundhs. Real-time epipolar geometry estimation and disparity. In *International Conference on Computer Vision, (Kerkyra, Greece)*, pages 234–141, 1999.

[Fletcher *et al.*, 2003] Luke Fletcher, Nicholas Apostoloff, Lars Petersson, and Alexander Zelinsky. Vision in and out of vehicles. *IEEE Intelligent Vehicles Symposium (IV2003)*, 2003.

[Fod *et al.*, 2002] Ajo Fod, Andrew Howard, and Maja J. Mataric. Laser-based people tracking. *In Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3024–3029, 2002.

[Kleeman, 1996] Lindsay Kleeman. *Understanding and Applying the Kalman Filter.* 1996.

[McCarthy and Barnes, 2004] C. McCarthy and N. Barnes. Performance of optical flow techniques for indoor navigation with a mobile robot. *Proceedings of IEEE International Conference on Robotics and Automation*, 2004.

[McKenna *et al.*, 2000] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80:42–56, 2000.

[Mittal and Davis, 2002] Anurag Mittal and Larry S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In *ECCV (1)*, pages 18–36, 2002.

[Porr *et al.*, 2002] Bernd Porr, Bernd Nrenberg, and Florentin Wrgtter. A vlsi-compatible computer vision algorithm for stereoscopic depth analysis in real-time. *International Journal of Computer Vision*, 49(1):39–55, August 2002.

[Saragih *et al.*, 2004] J. Saragih, D. Austin, and R. Goecke. Initialisation of model-based face tracking for a robot comedian. *Proceedings of Tenth Australian International Conference on Speech Science and Technology*, 2004.

[Sogo *et al.*, 2000] T. Sogo, H. Ishiguro, and M. Trivedi. N-ocular stereo for real-time human tracking, 2000.

[Sutherland *et al.*, 2000] Orson Sutherland, Harley Truong, Sebastien Rougeaux, and Alexander Zelinsky. Advancing active vision systems by improved design and control. *Proceedings of International Symposium on Experimental Robotics*, December 2000.

[Thompson, 2002] S. Thompson. *A Multi-Level Spatial Memory for Vision-Based Mobile Robot Localisation.* PhD thesis, Australian National University, 2002.

[Tomasi, 2000] Carlo Tomasi. *Mathematical Methods for Robotics and Vision.* 2000.

[Watman *et al.*, 2004] C. Watman, D. Austin, N. Barnes, G. Overett, and S. Thompson. Fast sum of absolute differences visual landmark detector. *Proceedings of IEEE Conference on Robotics and Automation*, April 2004.

[Welch and Bishop, 2004] Greg Welch and Gary Bishop. *An introduction to the Kalman Filter.* 2004.